Title

Accounting for undetected compounds in statistical analyses of mass spectrometry ‘omic studies

Permalink
https://escholarship.org/uc/item/9k18m8jz

Journal
Statistical Applications in Genetics and Molecular Biology, 12(6)

ISSN
2194-6302

Authors
Taylor, Sandra L
Leiserowitz, Gary S
Kim, Kyoungmi

Publication Date
2013-12-01

DOI
10.1515/sagmb-2013-0021

Peer reviewed

Published in final edited form as:
Stat Appl Genet Mol Biol. 2013 December 1; 12(6): 703–722. doi:10.1515/sagmb-2013-0021.

Accounting for Undetected Compounds in Statistical Analyses of Mass Spectrometry ‘Omic Studies

Sandra L. Taylor[1,*], **Gary S. Leiserowitz**[2], and **Kyoungmi Kim**[1]

[1]Division of Biostatistics, Department of Public Health Sciences, University of California School of Medicine, Davis, CA, USA

[2]Division of Gynecologic Oncology, UC Davis Medical Center, Sacramento, CA, USA

Abstract

Mass spectrometry is an important high-throughput technique for profiling small molecular compounds in biological samples and is widely used to identify potential diagnostic and prognostic compounds associated with disease. Commonly, this data generated by mass spectrometry has many missing values resulting when a compound is absent from a sample or is present but at a concentration below the detection limit. Several strategies are available for statistically analyzing data with missing values. The accelerated failure time (AFT) model assumes all missing values result from censoring below a detection limit. Under a mixture model, missing values can result from a combination of censoring and the absence of a compound. We compare power and estimation of a mixture model to an AFT model. Based on simulated data, we found the AFT model to have greater power to detect differences in means and point mass proportions between groups. However, the AFT model yielded biased estimates with the bias increasing as the proportion of observations in the point mass increased while estimates were unbiased with the mixture model except if all missing observations came from censoring. These findings suggest using the AFT model for hypothesis testing and mixture model for estimation. We demonstrated this approach through application to glycomics data of serum samples from women with ovarian cancer and matched controls.

Keywords

point-mass mixture; accelerated failure time model; missing values; metabolomics; glycomics; mass spectrometry

1. Introduction

Mass spectrometry has become an important analytical technique for profiling a wide array of compounds such as proteins, metabolites, lipids, and glycans in biological samples. This technology allows investigators to identify and quantify the suite of compounds present in biological samples and is now being widely used to identify compounds as potential diagnostic and prognostic tests, understand biological pathways of disease, and identify potential therapeutic targets. Raw data from a mass spectrometry analysis consists of the observed mass-to-charge ratios, the retention time if liquid or gas chromatography is used for separation prior to injection into the mass spectrometer and a measure of ion intensity (Enot et al. 2011). The mass-to-charge ratios and the retention times serve to identify unique compounds while the ion intensity yields a measure of each compound’s relative abundance.

[*]Corresponding author (ST): sltaylor@ucdavis.edu, phone: (916) 703-9171.

Extensive pre-processing of the raw data including baseline correction, noise reduction, smoothing, peak detection and alignment and peak integration is necessary before analysis (Want and Masson 2011). The final output of this processing is a data matrix consisting of the unique compounds identified and intensity measures of these compounds in each sample.

Depending on the specific technology used, the focus of the study (e.g., glycomics, proteomics, metabolomics), and the sample characteristics, hundreds to thousands of compounds can be identified. However, a common characteristic of the data from these studies is a large number of missing values. Hrydziuszko and Viant (2012) reported that 51.67% to 78.73% of the peaks in four metabolomics data sets contained missing values with the overall percentage of missing values ranging from 14.63% to 28.53%.Wang et al. (2012) note that commonly 20–40% of potential values are missing in quantitative mass spectrometry-based proteomics. In our glycomics studies, we have observed 55% to over 90% of the glycans to have missing observations for at least one sample and for the overall percentage of missing values to range from 20 to 80%. For individual glycans, the percentage of missing values can range from no missing values to more than 90% missing such as when a glycan is detected in only one sample.

Missing values can result from several mechanisms which might be influenced by the technology in use. A compound can be present in a sample but at a concentration below the detection limit of the mass spectrometer. Alternatively, a particular compound could be truly absent from a sample due to biological reasons. For example, a particular biological pathway could be suppressed in certain genetic variants such that compounds of this pathway are absent in these variants (Kliebenstein et al 2001; Kliebenstein et al. 2001a; Burow et al; 2010). Finally, a compound can be present in a sample at a level above the detection limit but fail to be detected due to technical issues related to sample preparation or processing (Bell 2009); Eidehammer et al. 2013; Michalski et al. 2011; Wang 2009). Regardless of the mechanism, in all three cases, a compound would be reported as a missing value in the resultant data set.

Missing data can be classified according to the properties of the processes causing the “missingness” (Little and Rubin 2002). Data are “missing completely at random (MCAR)” if the probability that an observation is missing is unrelated to its value and the values of other variables. Missing values resulting from technical measurement errors reflect an MCAR process because the value of the missing observations is independent of its value had it been observed and the value of other peaks. In contrast, for compounds that are censored, their missingness results because they occur at low abundances, below the detection limit of the instrumentation. Missingness for these compounds is “missing, not at random (MNAR)” because the probability that the observation is missing is related to the value of the missing observation. Hrydziuszko et al. (2011) found that missing values in four metabolomics data sets did not conform to MCAR. In fact, they found greater missingness with decreasing abundance suggestive of censoring. In proteomics data generated through liquid chromotography-Fourier transform ion cyclotron mass spectrometry,Karpievitch *et al.* (2009) also found a higher proportion of missing values among low-abundance peptides/proteins suggesting the origin of some missing values from detection limit censoring and thus MNAR. Hence, considering all potential causes, missingness of data generated by a mass-spectrometry likely occurs in a combination of MCAR and MNAR.

Several strategies are available for statistically analyzing data that contains missing values. These strategies can be broadly divided into methods that explicitly account for missing values in the modeling and those that manipulate the data to eliminate missing values prior to analysis. Strategies in the latter group include dropping samples or compounds with missing values or applying various imputation methods. A simple approach is to drop samples with missing values for a particular compound but this approach introduces bias to estimation and reduces the power of the test. In studies with small sample sizes, dropping samples with missing values could result in a prohibitively small sample size for analysis. More commonly, individual compounds with a large proportion of missing values are dropped from the analysis, with compounds observed in a sufficient number of samples retained in the analysis. Values for the remaining missing values are typically imputed and then standard statistical methods applied.

Many methods for imputing missing values are available and have been widely used in microarray studies (for review see Aittokalio 2009). In microarray studies, missing values have been considered to arise randomly from technical issues and have been treated as MCAR (Aittokalio 2009). In mass spectrometry ‘omics studies, however, missing values can originate from detection limit censoring and hence are MNAR. Because most imputation techniques produce unbiased results only if the missing data are MCAR or missing at random, but not MNAR (Karpievitch et al. 2012, Lee 2004), using the imputation methods developed for microarrary studies in mass spectrometry ‘omics studies could lead to biased results. Further, the choice of imputation method can substantially affect the results and interpretation of analyses of metabolomics data (Hrydziuszko and Viant 2012).

An alternative strategy is to use methods that explicitly account for missing values. Several statistical methods have been developed, each modeling the missing values in a different manner and based on different assumptions about the mechanism causing the missing values. Survival analysis methods (see e.g. Klein and Moeschberger 2003) can be used to model compounds that are present but censored at concentrations below the detection limit. Two-part models (Lachenbruch 1976, Lachenbruch 1992, Lachenbruch 2001, Duan et al. 1983, Taylor and Pollard 2009) are appropriate when missing values represent either the true absence of a compound or reflect technical measurement errors. Mixture models combine elements from survival analysis and two-part models.

Survival (i.e., time-to-event) data is characterized as being non-negative, and censoring which may be left, right or interval censoring, is common. Statistical methods specific for time-to-event data are necessary to properly model these unique data characteristics. Although not time-to-event data, mass spectrometry ‘omics data share the characteristics of being non-negative and possibly left-censored (Karpievitch et al. 2009). Tekwe et al. (2012) investigated using methods from survival analysis, specifically an accelerated failure time (AFT) model, to analyze proteomics data with missing values. Under the AFT model, missing observations are assumed to come from a continuous distribution but are below a detection limit and hence are not observed. With greater than 5% missing values, the AFT model had higher power to detect mean differences between experimental groups than standard two-sample parametric and non-parametric tests applied to the same data with values imputed for missing values using several methods (Tekwe et al 2012). The superiority of the AFT model increased as the percentage of missing values increased.

At the other end of the spectrum from survival analysis methods are two-part models. In these models, all missing values are modeled as a “point-mass”, typically at zero. A “point-mass” is a spike in the compound’s distribution resulting from the multiple missing observations at a specific value, in this case zero. The compound’s distribution is characterized by the proportion of observations in the point-mass representing the proportion of missing values and the distribution of the observed values. Two-part models consist of two parts, with one part testing for a difference in the proportion of missing values and another part that tests for a difference in the means of the continuous component. Unlike the AFT model, the missing values concentrated in the point-mass do not contribute to estimating the mean of the continuous component. Further, two-part models test the joint null hypothesis that both the proportions of missing values and the means of the continuous component are the same for the covariates of interest such as disease state or survival status. This hypothesis contrasts with the null of hypothesis of the AFT model and methods used for data assumed to come wholly from a continuous distribution such as *t*-tests (i.e., continuous data methods) which test for a difference in means between groups.

In the context of mass spectrometry metabolomics data, Taylor and Pollard (2009) compared several two-part models to standard continuous data two-sample tests. They showed that the two-part models were superior to standard two-sample tests when the group with the higher proportion of missing values also had a higher mean in the continuous component.Wood et al. (2004) used a two-part likelihood ratio test to analyze two-dimensional polyacrylamide gel electrophoresis proteomics data and also showed the two-part model to outperform a t-test when more missing values are present in the group with higher mean observed values. Similar to two-part tests,Wang et al. (2012) proposed a two-step approach to analyzing proteomics data. They suggested testing for differences in 1) the presence/absence of a peptide with an exact bionomial likelihood test and then 2) intensity measurements with a regression model.

Neither the two-part models nor survival analysis techniques fully capture the missing value processes present in mass spectrometry data. Mixture models combine aspects of the two-part and AFT models to account for missing values resulting from a combination of processes, including censoring, true absence and technical variability resulting in the failure to detect a compound (Moulton and Halsey 1995). These models can be used to test for covariate effects on both the proportion of observations in the point mass and on the mean of the continuous component, and further provide a means to test for the presence of a point mass versus origination of missing values only through censoring. Despite the potential utility of these models for ‘omics data, studies on their performance in this setting are largely absent.

Wu et al. (2009) considered application of a mixture model consisting of a binomial distribution and a truncated normal distribution to proteomics data generated by two-dimensional polyacrylamide gel electrophoresis to detect differences between two groups, comparing the power and Type I error rate of this model to *t*-tests through simulations. Interestingly, for many of the simulated data sets, *t*-tests with the simple imputation approach of substituting the global minimum for missing values, yielded similar or higher power to detect group mean differences than the mixture model. However, the mixture model outperformed the *t*-test when the group means differed but the proportion of missing values was the same between the two groups. In the context of mass-spectrometry proteomics data, Karpievitch et al. (2009) used simulations to compare a mixture model consisting of a binomial distribution and a log normal distribution to an ANOVA in which missing values were imputed using row-mean imputation. They found the mixture model to have higher sensitivity for a given specificity than the ANOVA to detect group mean differences with the mixture model performing relatively better as the proportion of missing values increased.

These previous studies have shown methods that explicitly account for missing values (AFT model, two-part model, mixture model) to frequently have better power than continuous data methods to detect differences in means between groups. However, no studies have compared the power (i.e., the probability of rejecting the null hypothesis of no difference between groups when it is false), and Type I error rates (i.e., the probability of rejecting the null hypothesis of no difference when in fact it is true) of mixture models and AFT models in the context of mass spectrometry data. Further, while detecting differentially regulated compounds is typically the primary interest in ‘omics studies, estimation of effect sizes and standard deviations is also of interest but has not been investigated for these models when missing values arise from multiple processes. Finally, the mixture and AFT models can be used to test for the presence of a point mass (Mouton and Halsey 1995); power and Type I error rates of this test have not been investigated.

Here, we compare performance of a mixture model to an AFT model. Section 2 presents the mathematical formulation of the two models and their respective hypothesis tests. In Section 3, we use simulations to compare 1) the Type I error rates and power of the tests under a range of conditions for the proportion and type of missing values (e.g., censored, true zeros, or randomly unobserved) and parameters for the continuous component, 2) parameter estimates from the two models and 3) the Type I error rates and power of tests for detecting a point mass at zero. This section culminates with our proposed two-step testing and estimation strategy for analyzing ‘omics data with missing values. We then apply our proposed strategy to liquid-chromotography mass-spectrometry glycomics data from serum samples of women with and without ovarian cancer (Section 4). Throughout the sections, in addition to comparing the mixture and AFT models, we also use a standard analytical approach of imputing missing data via *k*-nearest neighbor (KNN) and testing for group differences with a *t*-test. In Section 5, we discuss the implications of our findings and opportunities to further customize and extend the mixture model to a wide range of applications.

2. Model Formulation and Methodology

2.1 Accelerated Failure Time and Mixture Models

Consider first the AFT model. We assume a log-normal distribution because log transformed metabolomics data are usually approximately normally distributed (Karpievitch et al. 2012). Further, using a log-normal distribution is consistent with typical analytical approaches to ‘omics data that log transform intensity values and then use a *t*-test, ANOVA, or linear regression which assume normally distributed data. Although we use a log-normal distribution, many other parametric distributions can be used including gamma, Weibull, log-logistic, exponential, and inverse Gaussian (Klein and Moeschberger 2003). For any particular data set one of these distributions could better fit the data. Both the AFT and mixture models are flexible with respect to the parametric distribution.

In this AFT model, small molecular compound (e.g., protein, metabolite, or glycan) concentrations (*Y*) arise from a log-normal distribution with mean μ and standard deviation σ. A compound is observed if its concentration (*y*) is greater than or equal to the detection limit, *T*, of the analytical method, otherwise *y* is missing. In the AFT model, all missing values are assumed to result from censoring of concentrations below the detection limit. The likelihood function of this AFT model for a compound is

$$L(\mu_i,\sigma)=\prod_{i=1}^{n}\{\Phi([\mathit{log}(T)-\mu_i]/\sigma)\}\delta^{i}\cdot\{\mathit{exp}(-[\mathit{log}(y_i)-\mu_i]^{2}/2\sigma^{2})/y_i\sqrt{2\pi}\sigma\}^{1-\delta_i}$$

where *T* is the detection limit of the analytical method, σ is the standard deviation on the log scale, μ_i is the log scale population mean for covariate values of subject *i*, y_i is the observed compound intensity, δ_i is a censoring indicator that equals to 0 if $y_i \geq T$ and 1 otherwise, Φ is the cumulative standard normal distribution function, and *n* is a sample size of subjects.

Covariates are incorporated by setting $\mu_i=\boldsymbol{x}_i'\boldsymbol{\beta}$ where $\boldsymbol{x}_i'$ is a vector of covariate values for subject *i* and β is a vector of coefficients.

The AFT model assumes all missing values result from censoring. However, in a population, unobserved values of y can reflect a censored non-zero value (i.e., $0 < y < T$), the complete absence of the compound (i.e., $y = 0$), or the failure to observe y due to random technical issues even though y is present in the sample at levels greater than T. To account for individuals for which a compound is absent or unobserved for reasons other than censoring, another parameter, τ_i is added to the likelihood (Moulton and Halsey 1995). Let τ_i be the probability that for subject i, y_i emanates from a log-normal distribution. Then $1 -\tau_i$ is the probability that y_i is in a point mass at 0. The likelihood function for this mixture model for a compound is

$$L(\mu_i,\sigma)=\prod\nolimits_{i=1}^{n}\left\{(1-\tau_i)+\tau_i\cdot\Phi\left(\frac{[log(T)-\mu_i]}{\sigma}\right)\right\}^{\delta_i}\cdot\left\{\tau_i\cdot exp(-[log(y_i)-\mu_i]^2/2\sigma^2)/y_i\sqrt{2\pi}\sigma\right\}^{1-\delta_i} \quad (1)$$

In this model, all observed values ($\delta_i = 0$) arise from a log-normal distribution while missing values ($\delta_i = 1$) consist of censored values from the log-normal distribution, true zeros and randomly unobserved values. Covariate effects on μ_i are incorporated as in the AFT model. For covariate effects on τ_i, we use the logistic model (Moulton and Halsey 1995), specifically

$$\tau_i=exp(\boldsymbol{z}_i'\boldsymbol{\gamma})/(1+exp(\boldsymbol{z}_i'\boldsymbol{\gamma})) \quad (2)$$

where $\boldsymbol{z}_i'$ is a vector of covariate values for subject i and γ is a vector of coefficients.

2.2 Hypothesis Testing

The AFT and mixture models differ by the parameter, τ and the AFT model is nested within the mixture model. Thus, a likelihood ratio test can be used to compare these models and specifically test for whether the addition of a point mass at zero provides a better fit for the data than a censored only model (i.e., test the null hypothesis H_o: $\tau = 1$) (Moulton and Halsey 1995). Because the null hypothesis is on the boundary of the parameter space, the null distribution is a mixture of a $\chi_0^2 \text{and} \chi_1^2$, commonly assumed to be a 50:50 mixture (Self and Liang 1987).

Likelihood ratio tests also can be used with the mixture model to test the significance of effects of individual covariates on μ and/or τ. A single degree of freedom χ^2 likelihood ratio test can be used to evaluate the effect of one covariate on one parameter (μ or τ). Alternatively, a two degrees of freedom χ^2 likelihood ratio test can be used to test the effect of a covariate on both μ and τ.

2.3 Model Fitting

The AFT model is easily fit with standard statistical software. We used the survreg function in the survival package (Therneau and Grambsch 2000) for R version 2.15.1 (R Core Team 2012). In mass spectrometry studies, the detection limit for the machine is not fixed and known. Because parameters of the AFT model are fit using maximum likelihood estimation, we use the maximum likelihood estimate for the detection limit for each compound which is the minimum observed value.

To fit the mixture model, we largely followed Moulton and Halsey's approach to maximizing the likelihood. Starting values for β were obtained with a linear regression using log transformed observed values. For γ, we estimated the intercept value of τ and set the remaining γ's to 0 as suggested by Moulton and Halsey. We obtained an estimate for τ by first estimating the proportion of that would be below the minimum observed value for a compound assuming a log normal distribution with the mean and standard deviation of the observed values. This provided a crude estimate of the proportion of the population that was censored. We then estimated 1-τ as the difference between estimated censored proportion and the observed proportion of missing values. During development of this method, we investigated the sensitivity of the method to starting values of τ and found the optimization results to be robust to changes in starting values of τ.

With these starting values we used the quasi-Newton optimization routine of the optim function in R. This optimization approach occasionally failed to yield reasonable results for some simulated data sets. In these cases, we found the PORT optimization routines of the nlminb function to perform well. In both cases, the search space for the standard deviation was constrained to be positive but was unconstrained for the other parameters. R code for fitting the mixture model is provided in Appendix 2 and is available from the first author.

2.4 Imputation Followed by Application of *t*-test

A common approach to analyzing ‘omics data containing missing values is to impute the missing values to create a complete data set and then to apply a standard statistical method (see e.g. Eidhammer et al. 2013). For comparative purposes, we imputed missing values using *k*-nearest neighbors imputation with *k* set at 10. Imputation was accomplished with the R function impute.knn in the impute package (Hastie et al. 2012). Data were log transformed and group means compared with a *t*-test.

3 Simulation Study

3.1 Data Generation

Data sets consisting of two samples representing Controls and Cases were simulated assuming a range of parameters for the population means (μ), standard deviations (σ), and the proportion of values representing a point mass at zero (1-τ). Data were drawn from a log normal distribution. For each combination of simulation parameters, we generated 10,000 data sets. Each data set consisted of two groups (Cases and Controls) with sample sizes of 20, 30, 50 or 100. We then tested for differences between Cases and Controls for each of the 10,000 data sets. The power and Type I error rate of each method was calculated as the proportion of the 10,000 data sets with a significant p-value at the 0.05 level (raw p-value < 0.05).

For the simulations, the mean (log scale) of the Control group was fixed at 13. We note that the magnitude of intensity values in analytical data sets of ‘omics data are largely determined by the normalization procedure. In our work, we typically use a total quantity normalization procedure, scaling intensity values to the median total ion count across samples. With this procedure, we have observed median compound values ranging from about 7 to 18. More important for differential analyses is the magnitude of the difference in means (i.e., the effect size) relative to the standard deviation. We considered three log scale standard deviations (0.25, 0.5 and 1) and three effect sizes (Δ) of 0.25, 0.5 and 1, with standardized effect sizes (i.e, mean difference/standard deviation) ranging from 0.25 (Δ = 0.25 and σ = 1) to 2 (e.g., Δ = 1 and σ = 0.5). These standardized effect sizes encompass the range observed in the ovarian cancer data (see Figure 9).

In real data sets the total proportion of missing values is known, but the proportion of the distribution that are in the point mass (1-τ) is not known. Our approach to modeling mixture data was to fix the total proportion of missing values at 25%, 50% or 75% for the Control group and to set the proportion of the missing values in the point mass at 0%, 30%, 70% or 100% of the missing values. With 0% of the missing values in the point mass, all missing values originated from censoring (i.e., τ = 1) and with 100% of the missing values in the point mass, none resulted from censoring (i.e., $\tau = 0$) but rather were due to either true absence or technical failure to identify or quantitatively measure a compound present at levels above the detection limit.

We first simulated data for which all missing values came from censoring. To simulate this data, we set the censoring level for the Control group at 25%, 50%, and 75%. Based on the known parameters of the Control group we identified the threshold of the log normal distribution that would yield the desired level of censoring for the Control group. This threshold was used as a detection limit and simulated values in both the Control and Case groups below this threshold were set to 0. This approach represented a realistic setting with a defined detection limit. Because the Case group always was set to have a higher mean, the proportion of missing values was smaller than in the Control group, ranging from 4.7 to 66.4% (Table 1). The proportion of missing values considered encompassed a wide range and covered most of the range observed in real data sets. For the glycomics data analyzed in Section 5, the proportion of missing values ranged from 0 to 94% for controls and 1.9 to 94% for cancer samples with median values of 59.3% [IQR: 28.8, 88.5] and 75% [IQR: 40.4, 86.5], respectively. With this simulation approach, the two groups differed in the proportion of missing values and the means of the continuous components.

We then simulated data for which the missing values consisted of a combination of censored values and point mass values. Here we fixed the total proportion missing at the values for the all censored simulations. Thus, for the Control group the total proportion of missing values was 25%, 50%, and 75% but for the Case group, the total proportion missing depended on the mean and standard deviation of the particular simulation (Table 1). Values of τ for each simulation were calculated based on the total proportion missing and the proportion of the missing values in the point mass. Given the range of proportions, means and standard deviations, the simulations encompassed a wide range of values for τ. For the various simulation scenarios, Supplemental Tables 1 to 4 show the total proportion of missing values for the Control and Cases group and the proportion of the distribution in the point mass for each group. With this simulation approach, μ and τ differed between Cases and Controls for all the data sets.

3.2 Comparison of Type I Error Rates

We first compared the Type I error rates of the three methods by generating data for which the null hypothesis of no difference in μ and τ was true. All models controlled the Type I error rate at or close to the nominal level of 5% for all parameter combinations although, the mixture model was slightly conservative when all missing values were censored (i.e., 0% in the point mass) and the proportion of missing values was small (e.g. 25%). Conversely, the mixture model was slightly anti-conservative for some simulations with large proportions of missing values (e.g., 75%) and where most of the missing values were in the point mass. The AFT model maintained a narrow range of Type I error rates as did *t*test conjugated with KNN imputation. Table 2 shows the Type I error rates for data generated from a log normal distribution with $\mu = 13$ and $\sigma = 0.5$ on the log scale for group sample sizes of 30 and 50 for a range of missing values and point mass proportions. Results for other sample sizes and parameter combinations were similar (data not shown).

3.3 Comparison of Power to Detect Group Differences

Because all methods adequately controlled the Type I error rate, we directly compared the power of the models to detect differences between two groups for a range of sample sizes, standard deviations, group means, missing value proportions and point mass proportions. The AFT model consistently had equal or slightly higher power to detect differences between the two groups than the mixture model (Figures 1 & 2), yielding power up to 15%

higher. While we expected the AFT model to be superior to the mixture model when all missing values resulted from censoring, the higher power of the AFT model relative to the mixture model when a large portion of the missing values were in the point mass was unexpected. When the difference in the group means was twice as large as the standard deviation (e.g., σ = 0.5 and Δ = 1), both the mixture and AFT models had approximately 100% power to detect group differences at a 5% significance level even with a sample size of only 20 per group (data not shown). Power remained high for both methods when Δ was the same as the standard deviation, for sample sizes of 30 or greater. With a difference in means only half the size of the standard deviation, sample sizes greater than 50 were necessary to achieve 80% power.

Imputation with *t*-tests had the lowest power in most simulations. An exception was for simulations with 100% of the missing values in the point mass. In this case, imputation with *t*-test had the highest power. As discussed in Section 3.5, when 100% of the missing values are in the point mass rather than derived from censoring, estimates based on imputed data are less biased than when some of the missing values arise from censoring. This case is reflective of data that is only MCAR, a condition under which imputation with *t*-tests perform well.

The power of all methods declined as the total proportion of missing values increased. With 75% missing values, power was typically 10–20% lower than under similar simulation conditions but with only 25% missing values. The overall proportion of missing values in the point mass did not have a large effect on power of the AFT or mixture model although power of the mixture model increased slightly as the proportion of observations in the point mass increased.

Regardless of the method used, the presence of missing values reduces the power to detect group differences. For comparison, we calculated the power of a *t*-test to detect a mean difference of 0.25 for data with a standard deviation of 0.5 and samples sizes of 30 per group. Under these conditions, with no missing values a t-test yields 61% power. Assuming only censored values, the AFT model had 50% power with 25% missing values but only 35% power when 75% of the values were missing. Thus, even with using the AFT model which we found to have the highest power of the methods considered, power remains below what could be achieved with complete data.

3.4 Power to Detect Point Mass

For any given real data set, whether missing values reflect true zeros, censored values or values unobserved due to technical issues is unknown. The presence of a point mass at zero can be tested for using a likelihood ratio test to assess whether the censored (AFT) model or a model containing a point mass at zero in addition to censored values fits the data better. We considered the Type I error rate of this test by applying the test to data simulated with all missing values originating from censoring and then assessed power by applying it to data sets containing varying proportions of missing values in the point mass. Several patterns were apparent from these analyses.

First, because the parameter space is on the boundary, the true null distribution for the test statistic lies between 50:50 mixture of $\chi_0^2 \text{and} \chi_1^2$ distributions (Self and Laird 1987). To bound the Type I error rate, we evaluated the error rate under a χ_1^2 null distribution and a 50:50 mixture of $\chi_0^2 \text{and} \chi_1^2$ distributions. Clearly, the estimated Type I error rate was lower with χ_1^2 null distribution than with the 50:50 mixture null distribution given the lower significance threshold with the 50:50 mixture. Interestingly though, the 50:50 mixture null distribution did not control the error rate at the nominal level while the χ_1^2 null distribution was effective at maintaining the error rate close to the nominal level (Figure 3).

Because of the different Type I error rates depending on the null distribution used, we used Receiver Operating Curves to characterize and evaluate performance of the point mass test. Test performance was most strongly influenced by the proportion of the missing values in the point mass, the sample size and the total proportion of missing values. When all missing values were in the point mass, the test had high discriminatory ability to detect the point mass with area under the curve (AUC) values of 0.94 or greater (Figure 4). As the proportion of missing values in the point mass declined, test performance declined and was poor (AUC values less than 0.7) when only 30% of missing values were in the point mass even with a sample size of 100 (Figures 4). The total percentage of missing values was also influential, reducing the discriminatory ability of the test at high levels of missing values. The Type I error rate tended to increase as the total percentage of missing values increased, thereby reducing the test's overall performance.

3.5 Comparison of Parameter Estimates

Finally, we compared parameters estimates for the means of the log-normal distribution for each group and for the standard deviation under the range of simulated data conditions. The mixture model was superior to the AFT model and t-test with KNN imputation in estimating the distributional parameters. Estimates for the means and standard deviation from the mixture model were unbiased or nearly unbiased over much of range of standard deviations, mean differences, missing values and point mass proportions (Figures 5–7). The one exception to this superiority was if all missing values arose from censoring. In this case, estimates with the mixture model were slightly biased and less efficient that those with the AFT model (Supplemental Table 5). These characteristics worsened for the mixture model with increasing proportion of missing values while the AFT model remained unbiased.

In contrast when missing values consisted of censored values and point mass values, estimates from the AFT model were strongly affected by the total proportion of missing values and the proportion of missing values in the point mass. As the total and point mass proportions increased, the bias of the parameter estimates from the AFT model increased. With the AFT model, control means were generally under-estimated (Figure 5) and the standard deviations and effect sizes generally over-estimated with the AFT model (Figure 7). The bias increased with the total proportion of missing values and efficiency decreased.

Estimates from when KNN was used to impute missing values were biased in the opposite directions. Means were generally overestimated with the imputed data sets, and effect sizes and standard deviations were underestimated. Bias increased as the total proportion of missing values increased but declined as the proportion of observations in the point mass increased. In simulations with 100% of the missing values in the point mass, no values arise from censoring and the observed values reflect the underlying continuous distribution. In contrast when some of the missing values arise from censoring, none of the observed values come from lower tail of the distribution and parameter estimates are more biased than when a larger portion of the missing values are in the point mass.

4. Hypothesis Testing and Estimation Strategy for Point Mass Mixtures

The preceding results showed the AFT model to be more powerful for detecting differences between two groups with different means and point mass proportions than the mixture model and in most cases imputation with *t*-test. However, except when all missing values originated from censoring, the mixture model yielded better estimates of the means and standard deviations than the AFT model, particularly for data sets with a large proportion of missing values in the point mass. Based on these results, we suggest the following strategy for analyzing mass-spectrometry based ‘omics data with missing values. First, because of the greater power of the AFT model, we suggest testing for group differences with this model. Then, for compounds found to differ significantly between groups, fit the mixture model and conduct the point mass test to determine if including the point mass term yields a better fit to the data than the fully censored model. For compounds fit best with the mixture model, this model would then be used to estimate the population parameters.

5. Application to Ovarian Cancer Biomarker Project

5.1 Mass-spectrometry Glycomics Data

Liquid-chromotography mass spectrometry was used to analyze the glycome of serum of 52 women with stage III–IV ovarian cancer and 52 age-matched healthy controls. Out of 331 glycans potentially present, nine were not detected in any sample; only 32 glycans were detected in all samples. For this evaluation, we excluded glycans detected in all samples because data with no missing values are appropriately evaluated with standard methods. We also dropped 109 glycans detected in fewer than three samples in either group. These exclusions left 181 glycans for analysis.

5.2 Analysis Results

Of the 181 glycans analyzed, the mixture model identified the most (43) as significantly differing (raw pvalue < 0.05) by cancer status, followed by the AFT model (38) (Figure 8). Imputation with *t*-tests had the fewest significantly differing glycans (30). No adjustment for multiple testing was made because the goal of this analysis was to demonstrate a method rather than detect differentially-regulated compounds. The lower number of significant glycans detected through imputation with t-test is consistent with the simulation findings. Interestingly, although the simulations showed the AFT model to have higher power than the mixture model, the mixture model identified a few more glycans as significantly different between cancer groups. Most of the glycans that were significant with the mixture model but not the AFT model had very large proportions of missing values (>85%) and large observed effect sizes (Figure 9). We did not simulate data with extreme levels of missingness (>75%) and potentially the mixture model could outperform the AFT model under these circumstances. However, while we analyzed glycans with extreme levels of missingness for this example, retaining such compounds in practice may not be warranted due to the small number of observations on which to base inferences and estimations. Glycans found to differ significantly only with imputation and *t*-test approach also tended to have large proportions of missing values but with small observed effect sizes. For these glycans, the standard deviation was likely underestimated which then resulted in a significant difference despite a small effect size.

We then tested whether including a point mass at zero fit the data better than the fully censored model. For all significant glycans, the point mass test indicated that the mixture model fit the data better than the fully censored model (raw p-values < 0.05). Because the mixture model was deemed to fit the data better, we estimated the means and standard deviations for the continuous component using the mixture model and compared them to estimates from the AFT model. Consistent with the simulation results, mean estimates from the mixture model were higher on average than the AFT model and were less variable (Figure 10). The standard deviation estimates from the mixture model were smaller and similarly were less variable than the AFT model. Although the true means and standard deviations are unknown for these data, given the results from the simulation study, we expect the estimates from the mixture model to be less biased estimates than the AFT model.

6. Discussion

Mass spectrometry data for ‘omics studies typically have a large number of missing values in the resultant data sets. Missing values can result when a compound is present but at a concentration less than the detection limit of the analytical method, when the compound is completely absent from a sample, or when the compound is present in the sample at levels above the detection limit but is not observed due to technical issues. Imputation methods have limitations to truly reflect the nature of missingness consisting of both MCAR and MNAR processes present in most mass spectrometry data. Several previous studies have compared statistical methods for analyzing this type of data that explicitly account for missing values (e.g., AFT, two-part and mixture models) with standard complete data methods (e.g., *t*-test, Wilcoxon test, ANOVA). These studies found methods that specifically model missing values had higher power to detect group differences under some, but not all situations. The assumptions of the both the AFT and mixture model are reasonable for mass spectrometry data. Under the AFT model all missing values are assumed to result from censoring while the mixture model allows some missing values to be in a point mass at 0. The performance of these two models for analyzing mass spectrometry has, however, not been directly compared.

Through simulations, we found the AFT model to have greater power than the mixture model for detecting differences between two groups with different means and point mass proportions. However, the mixture model yielded unbiased or nearly unbiased and more efficient parameter estimates for the means and standard deviations while parameter estimates from the AFT model were biased when some of the missing values were in a point mass at zero rather than censored observations. Accurate parameter estimation is important for understanding the magnitude of the effect of the covariates and the variability of the compounds. Because of the greater power of the AFT model to detect differences between groups but the more accurate parameter estimates of the mixture model, we suggested using the AFT model to identify differentially-regulated compounds and then to estimate the parameters with the mixture model if the point mass test is significant.

In a real data set, for a given compound the mechanisms resulting in the missing values is unknown and could consist of a combination of detection limit censoring, technical failure to detect a compound and true absence of a compound. Because the missing value mechanism is unknown, we cannot determine whether the AFT or mixture model is the correct model and therefore risk mis-specifying the model. We showed that the AFT model generally provided equal or greater power than the mixture model regardless of whether the data arose from a mixture distribution or censored-only distribution. Thus, the impact of model mis-specification would be most evident in the parameter estimates. When the data came from a mixture distribution, the mixture model had largely unbiased estimates while the AFT model yielded highly biased and inefficient estimates when a large proportion of the missing values were in the point mass. In contrast, when all values resulted from censoring, estimates from the mixture model were only mildly biased and slightly less efficient than the AFT model estimates. We suggested using a combination of both models to take advantage of the strengths of each model while minimizing the impact of their respective short-comings.

In addition to parameter estimation, the mixture model is useful for testing whether the data are best fit by including a point mass at zero versus the AFT model which considers all missing values to originate from censoring. The simulations showed this test to have high power if all observations were in the point mass but little power if a relatively small proportion of the missing values were in the point mass. Because of the rather low power except for the most extreme circumstances, a more liberal significance level such as 0.2

could be appropriate in applying this test. Falsely rejecting the null hypothesis of no point mass, and using the mixture model to estimate parameters would not lead to poor estimates. However, failing to reject the null and using the AFT model for estimation could lead to biased estimates as shown in the simulations.

The mixture model is very flexible. In this study, for the mixture model, we assumed a log normal distribution for the continuous observations but other distributions, including the log gamma (Moulton and Halsey 1996), gamma, log-skew normal (Chai and Bailey 2008), and Weibull are also readily applicable where appropriate. Further, several other link functions, such as a probit (Duan et al. 1983) or complementary log-log link can be used to relate τ to the covariates. Both the parametric distribution and the link function can be changed as appropriate to fit to a particular data set.

The mixture model also provides a very flexible structure for hypothesis testing. First, unlike previous investigations that proposed and evaluated models for two or few groups (Wu et al. 2009, Karpievitch et al. 2009) our model formulation allows many covariates to be modeled. Second, the covariates used to model μ and τ can be entirely the same, different or a subset of each other. Thus, the mixture model allows testing for the effect of the covariates separately on μ and τ. The AFT model implicitly assumes that all missing values originate from censoring and this model evaluates the effect of covariates on mean levels. Potentially though, some missing values could reflect compounds that are absent or suppressed and the presence/absence of a compound could be influenced by different covariates than the magnitude of a compound once observed. For example, different pathways could be responsible for determining the presence/absence of compound than those that determine the level of a compound given that it is produced at all. The flexible structure of the mixture model provides a potential advantage over the AFT model in being able to test for these types of relationships and enhance analysis and interpretation of ‘omics data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We sincerely thank Drs. Renee Ruhaak and Carlito Lebrilla for their dedicated work in generating the ovarian cancer glycomics data. The project described was supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), through grant #UL1 TR000002. This work was supported by NIH/NIA grant P01AG025532 and the Ovarian Cancer Research Fund.

References

1. Burow M, Halkier BA, Kliebenstein DJ. Regulatory networks of glucosinolates shape *Arabidopsis thaliana* fitness. Current Opinion in Plant Biology. 2010; 13:348–353. [PubMed: 20226722]
2. Chai HS, Bailey KR. Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero. Stat. Med. 2008; 27:3643–3655. [PubMed: 18186536]
3. Duan N, Manning WG Jr, Morris CN, Newhouse JP. A comparison of alternative models for the demand for medical care. J. Bus. Econ. Stat. 1983; 1:115–126.
4. Enot DP, Haas B, Weinberger KM. Bioinformatices for mass-spectrometry-based metabolomics. Method Mol. Biol. 2011; 719:351–375.
5. Hastie T, Tibshirani R, Narasi B, et al. Impute: Imputation for microarray data. R package version 1.32.9. 2012
6. Hrydziuszko O, Viant MR. Missing values in mass spectrometry based metabolomics, an undervalued step in the data processing pipeline. Metabolomics. 2012; 8:161–174.

7. Karpievitch Y, Stanley J, Taverner T, et al. A statistical framework for protein quantitation in bottom-up MS-based proteomics. Bioinformatics. 2009; 25:2028–2034. [PubMed: 19535538]

8. Karpievitch Y, Dabney AR, Smith RD. Normalization and missing value imputation for label-free LC-MS analysis. BMC Bioinformatics. 2012; 13(Suppl 16):55. [PubMed: 22480135]

9. Klein, JP.; Moeschberger, ML. Survival Analysis: Techniques for Censored and Truncated Data. 2nd edition. New York: Springer-Verlag; 2003.

10. Lachenbruch PA. Analysis of data with clumping at zero. Biometrische Zeitschrift. 1976; 18:351–356.

11. Lachenbruch, PA. Utility of logistic regression analysis in epidemiologic studies of the elderly. In: Wallace, RB.; RF, Woolson, editors. Epidemiologic Methods in the Study of Aging. New York: Oxford University Press; 1992. p. 371-381.

12. Lachenbruch PA. Comparisons of two-part models with competitors. Statistics in Medicine. 2001; 20:1215–1234. [PubMed: 11304737]

13. Lee, ML. Analysis of microarray gene expression data. New York: Kluwer Academic Publishers; 2004.

14. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2nd Edition. Hoboken: John Wiley & Sons; 2002.

15. Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. Journal of Proteome Research. 2011; 10:1785–1793. [PubMed: 21309581]

16. Moulton LH, Halsey NA. A mixture model with detection limits for regression analyses of antibody response to vaccine. Biometrics. 1995; 51:1570–1578. [PubMed: 8589241]

17. Moulton LH, Halsey NA. A mixed gamma model for regression analyses of quantitative assay data. Vaccines. 1996; 14:1154–1158.

18. R Core Team. R, A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 2012. URL http://www.R-project.org/.

19. Self SG, Liang K-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. 1987; 82:605–610.

20. Taylor S, Pollard K. Hypothesis tests for point-mass mixture data with application to ‘omics data with many zero values. Stat. Appl. Genet. Mo. B. 2009; 8(1):1–43. Article 8.

21. Tekwe CD, Carroll RJ, Dabney AR. Application of survival analysis methodology to the quantitative analysis of LC-MS proteomic data. Bioinformatics. 2012; 28:1998–2003. [PubMed: 22628520]

22. Therneau T, Grambsch PM. Modeling Survival Data: Extending the Cox Model. Springer, N.Y. ISBN 0-387-98784-3. 2000

23. Wang X, Anderson GA, Smith RD, et al. A hybrid approach to protein differential expression in mass spectrometry-based proteomics. Bioinformatics. 2012; 28:1586–1591. [PubMed: 22522136]

24. Want E, Masson P. Processing and analysis of GC/LC-MS-Based metabolomic data. Method Mol. Biol. 2011; 708:277–298.

25. Wood J, White IR, Cutler P. A likelihood-based approach to defining statistical significance in proteomic analysis where missing data cannot be disregarded. Signal Process. 2004; 84:1777–1788.

26. Wu S, Black MA, North RA, et al. A statistical model to identify differentially expressed proteins in 2D PAGE Gels. PLOS Comp. Biol. 2009; 5(9):e1000509.

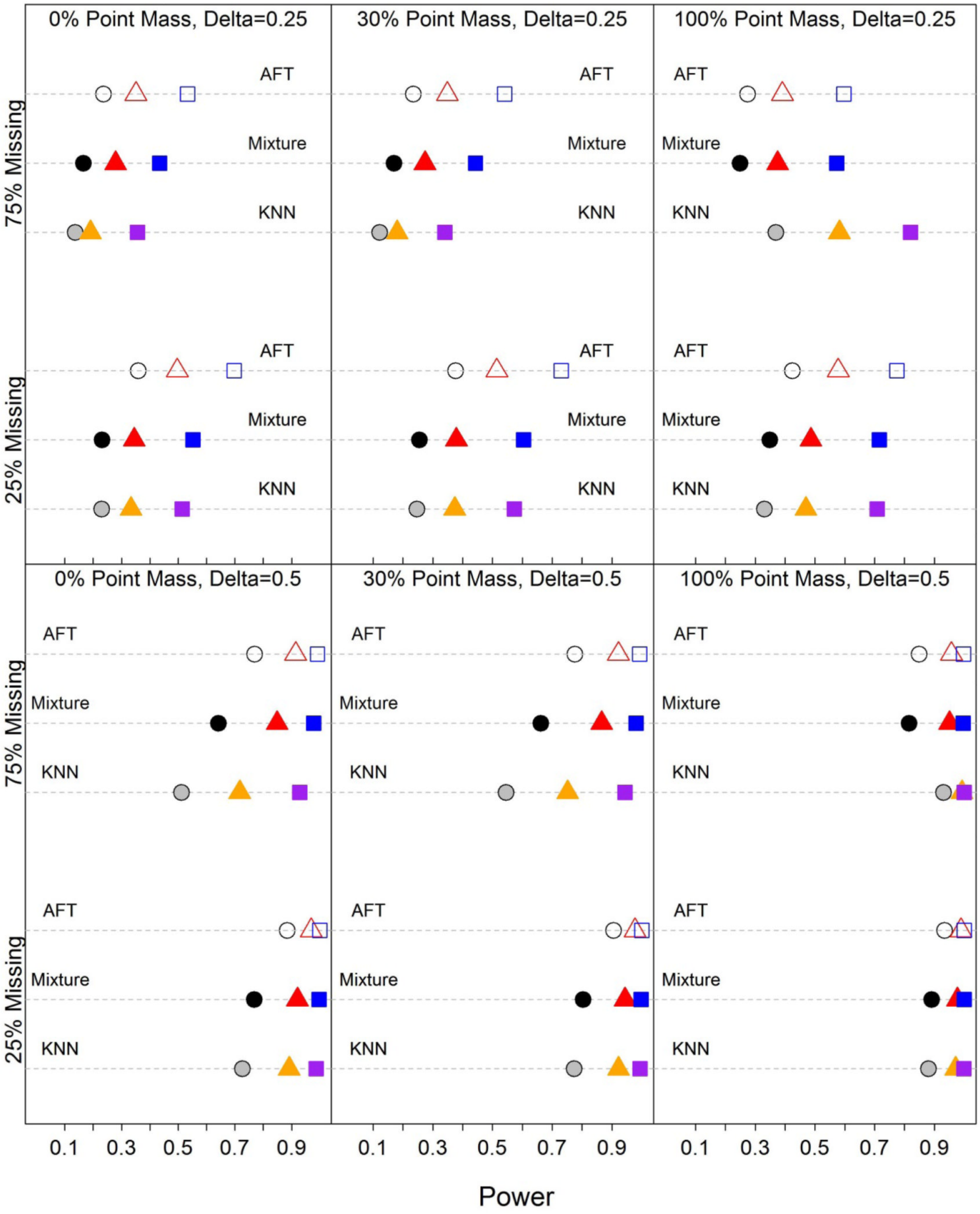

Figure 1.

Power of accelerated failure time (AFT) model and mixture model to detect differences between two groups with mean differences of Δ = 0.25, 0.5, and 1, total proportion of missing values of 25% and 75% and with the proportion of missing values in a point mass at 0 of 0%, 30%, 70% and 100%. The standard deviation of both groups was 0.5. Sample sizes in each group were 20 (●,○,●), 30 (▲, △, ▲) and 50 (■, □, ■).

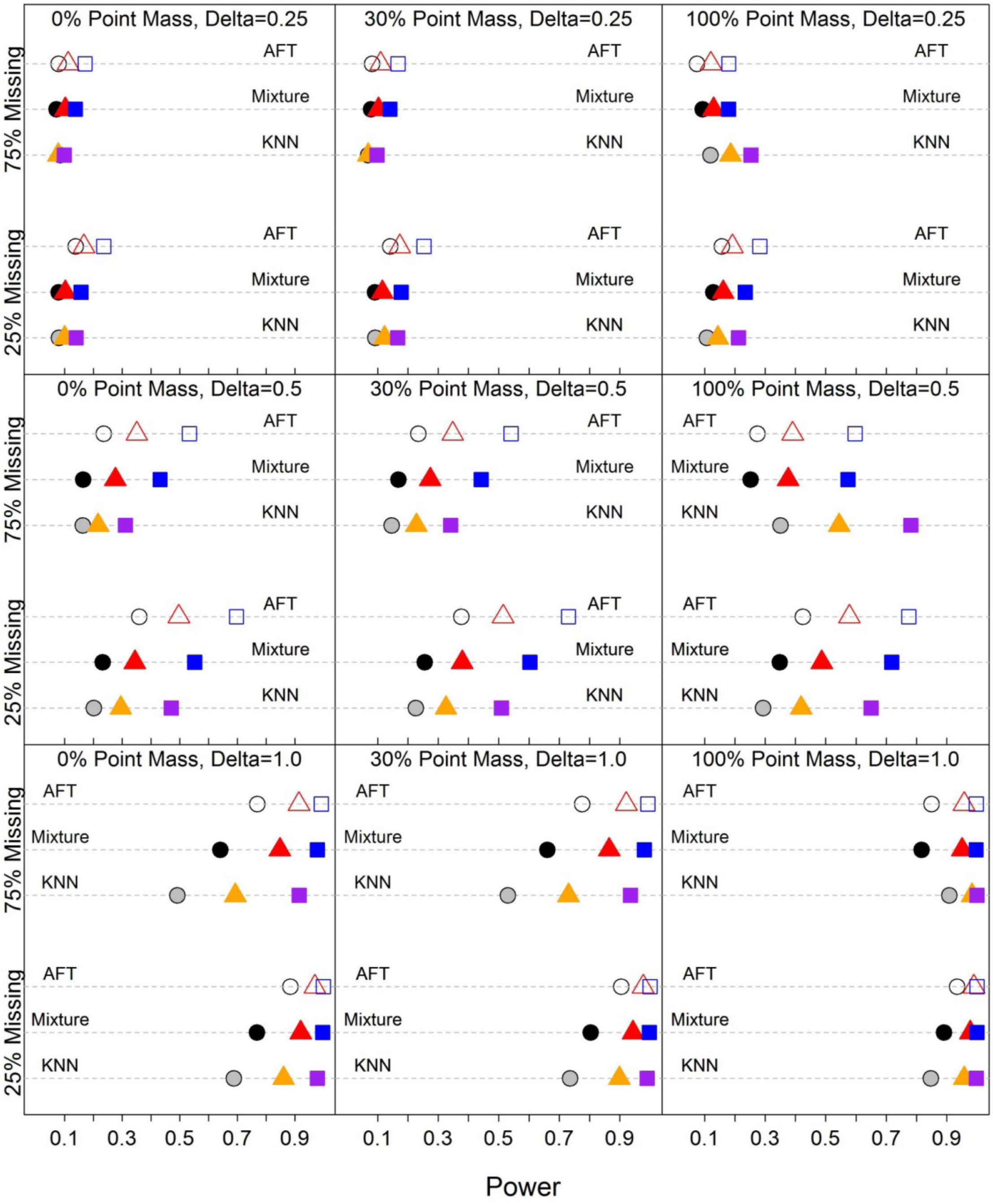

Figure 2.

Power of accelerated failure time (AFT) model and mixture model to detect differences between two groups with mean differences of Δ = 0.25, 0.5, and 1, total proportion of missing values of 25% and 75% and with the proportion of missing values in a point mass at 0 of 0%, 30%, 70% and 100%. The standard deviation of both groups was 1. Sample sizes in each group were 20 (●, ○, ●), 30 (▲, △, ▲) and 50 (■, □, ■).

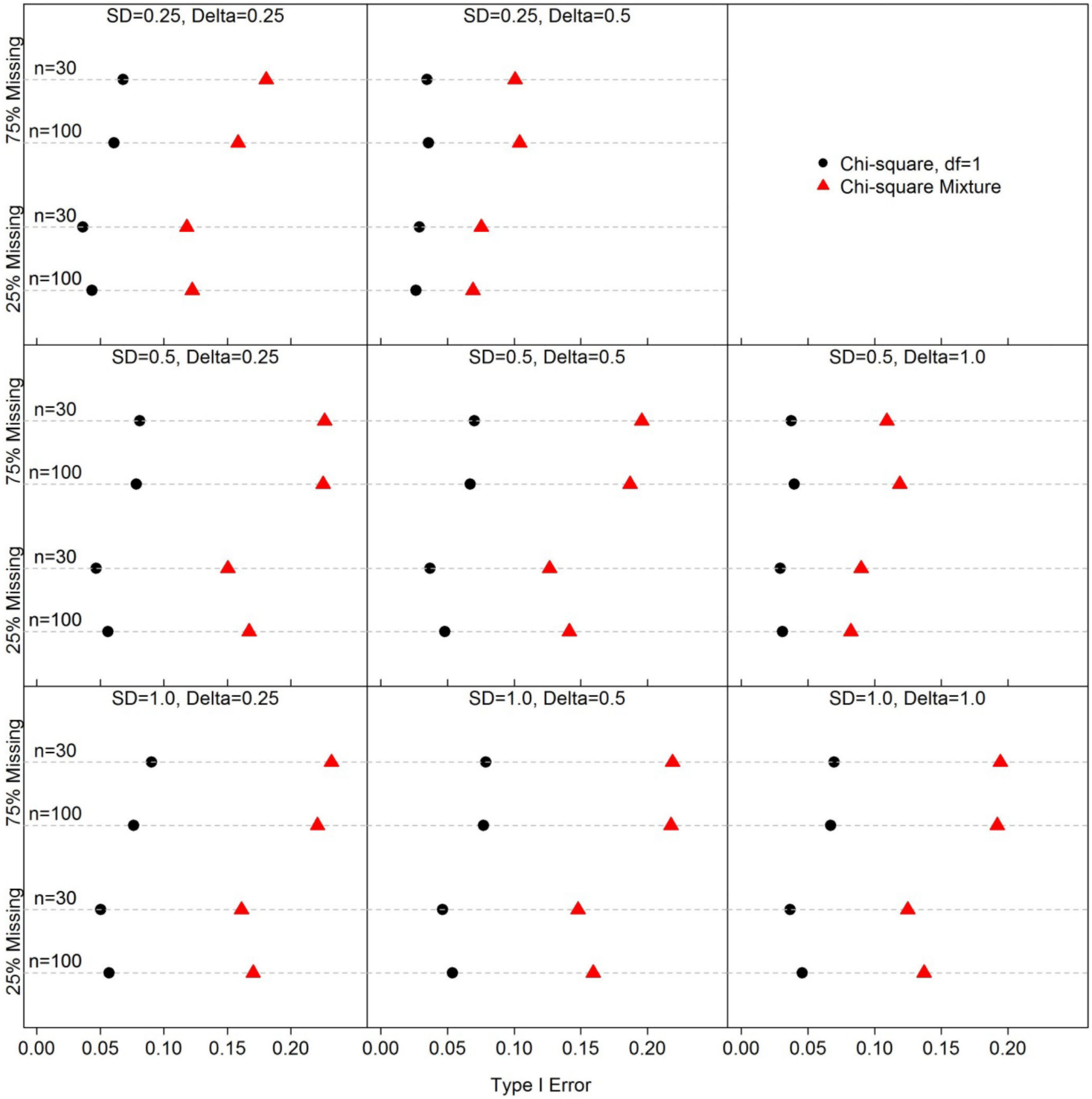

Figure 3.

Type I error of the point mass test for group sample sizes of 30 and 100 with varying proportions of missing values (25% and 75%), mean differences (Δ = 0.25, 0.5, or 1) and standard deviations (SD = 0.25, 0.5, and 1). Error rates shown are based on a χ_1^2 null distribution (●) and 50:50 mixture of $\chi_0^2 \text{and} \chi_1^2$ distributions as the null distribution (▲).

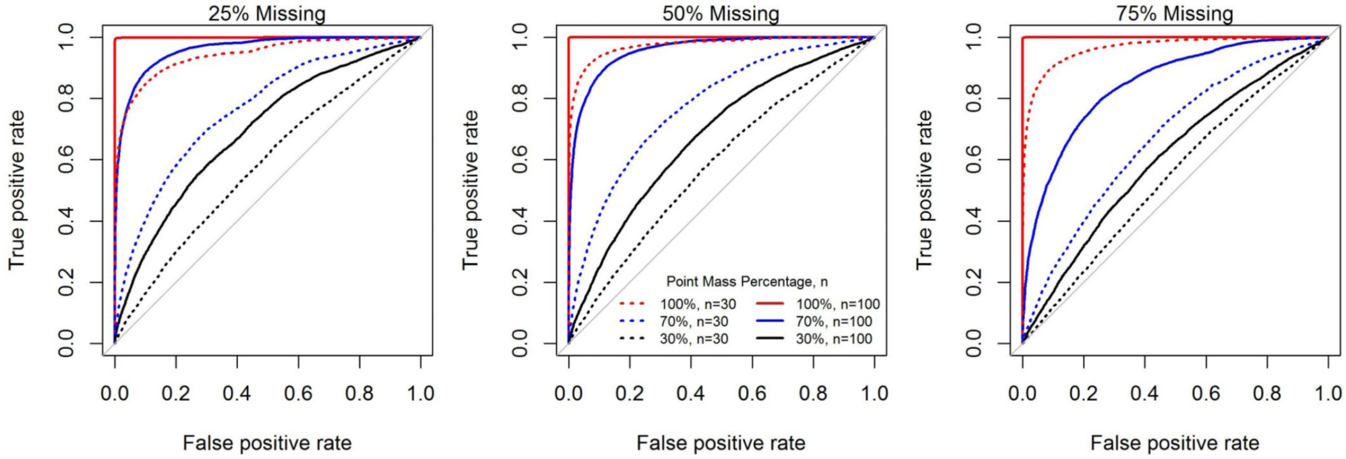

Figure 4.

Receiver-Operative Characteristic curves of the performance of the point mass test to detect a point mass for group sample sizes of 30 and 100 with varying proportions of missing values (25%, 50%, and 75%) and varying proportions in the point mass (30%, 70% and 100%). The mean differences between groups (Δ) was 0.5 and the standard deviation was 0.5.

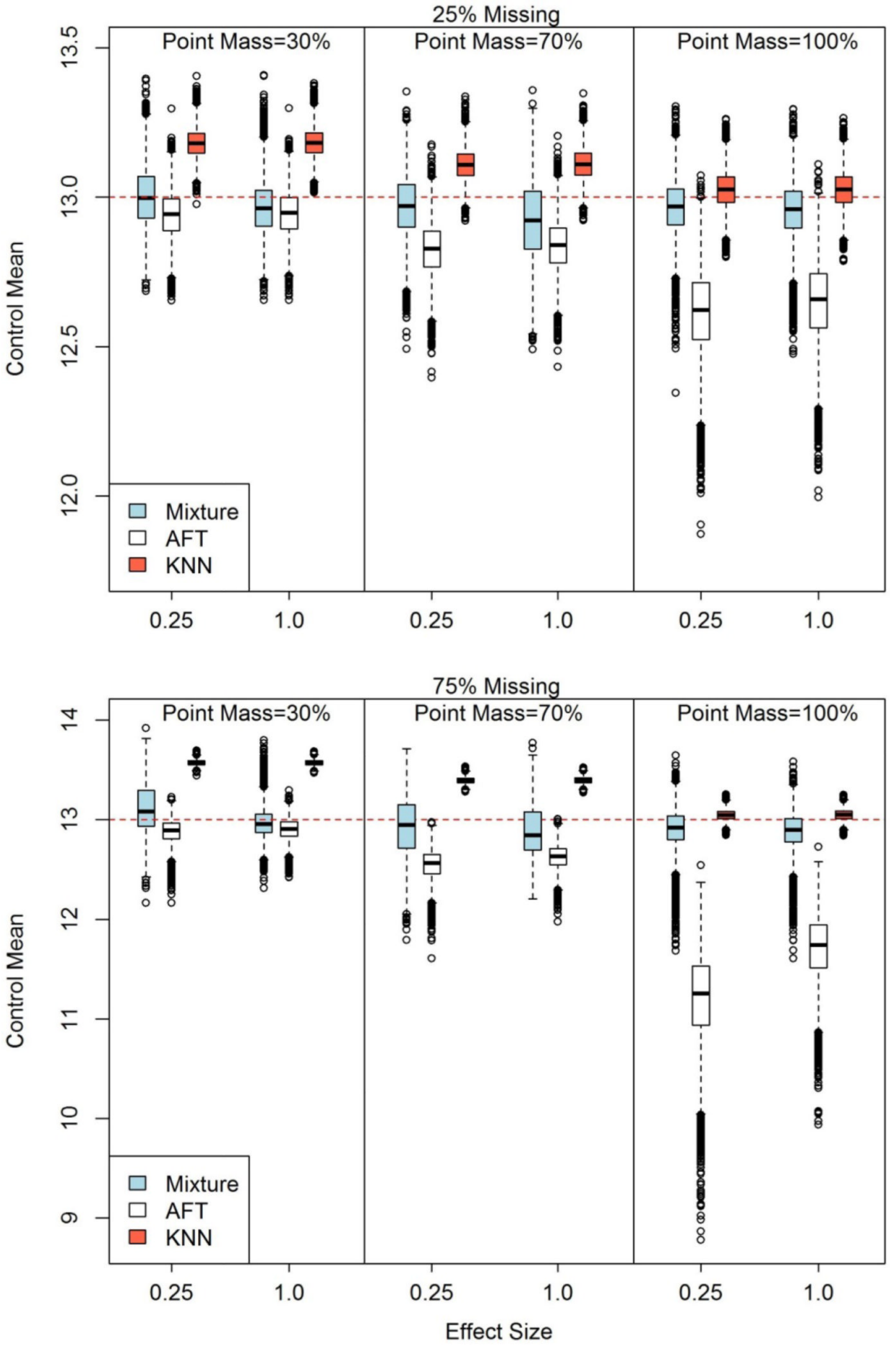

Figure 5.

Distribution of the mean estimates of the Control group based on the mixture, AFT, and KNN methods for 10,000 simulated data sets. Data were simulated from a log normal distribution with a mean of 13 for the Control group and means of 13.25 and 14.0 for the Case group reflecting effect sizes (i.e., Δ) of 0.25, and 1.0. The standard deviation was 0.5 for both the Control and Case groups. The proportion of missing values was 25 or 75% and the proportions of observations in point mass at 0 were 30%, 70%, or 100%. The sample size was 50 in each group. The horizontal red dashed line indicates the real value of the mean (13) in simulating data for the Control group. The effect size (x-axis) indicates the mean of the Case group relative to the Control.

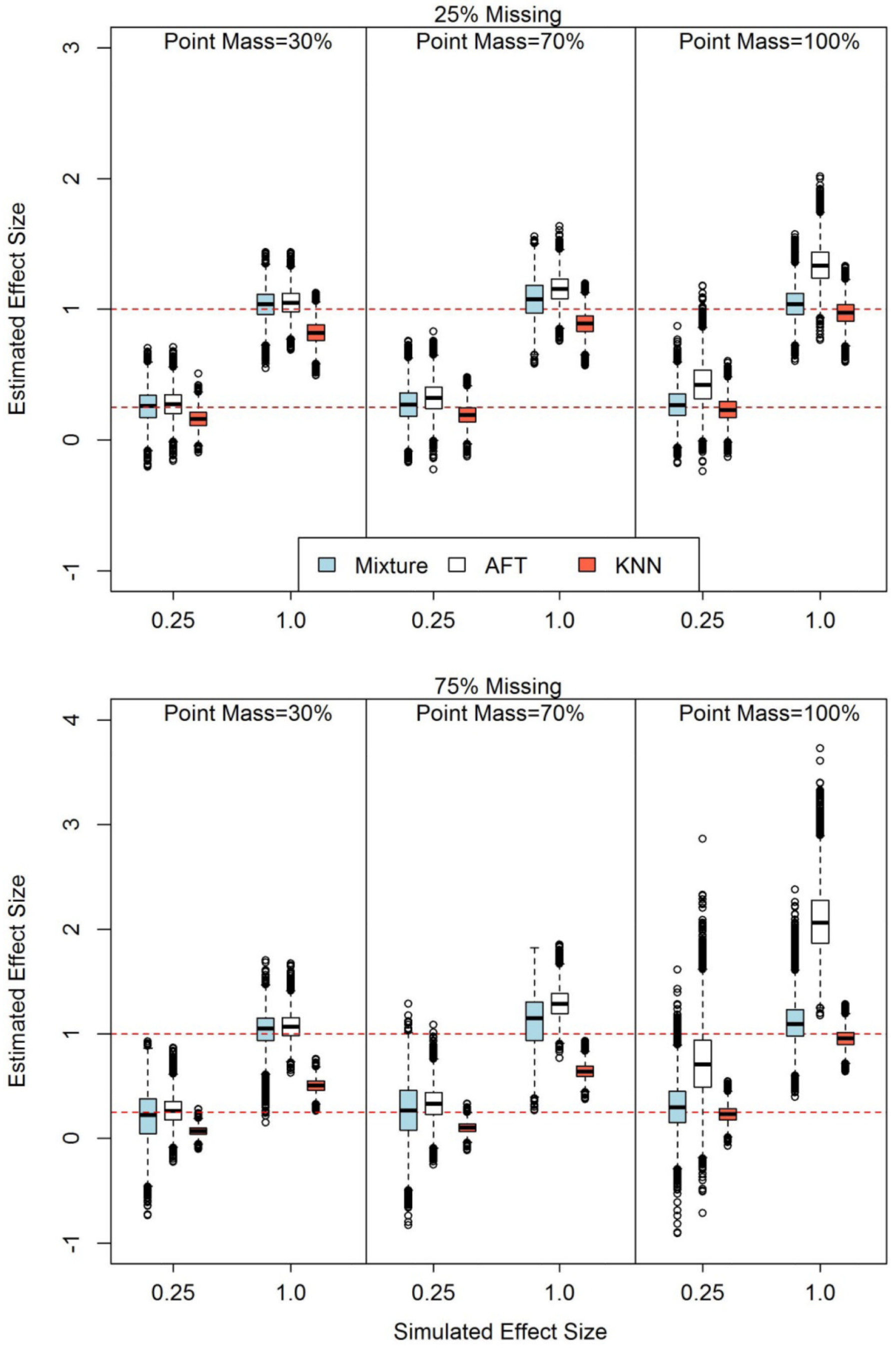

Figure 6.

Distribution of the effect size estimates (Δ) based on the mixture, AFT and KNN methods for 10,000 simulated data sets. Data were simulated from a log normal distribution with a mean of 13 for the Control group and means of 13.25 and 14.0 for the Case group reflecting effect sizes (i.e., Δ) of 0.25 and 1.0. The standard deviation was 0.5 for both the Control and Case groups. The proportion of missing values was 25 or 75% and the proportions of observations in point mass at 0 were 30%, 70%, or 100%. The sample size was 50 in each group. The horizontal red dashed lines indicate estimated effect sizes of 0.25 and 1. The x-axis shows the effect size used in the simulations (0.25 or 1).

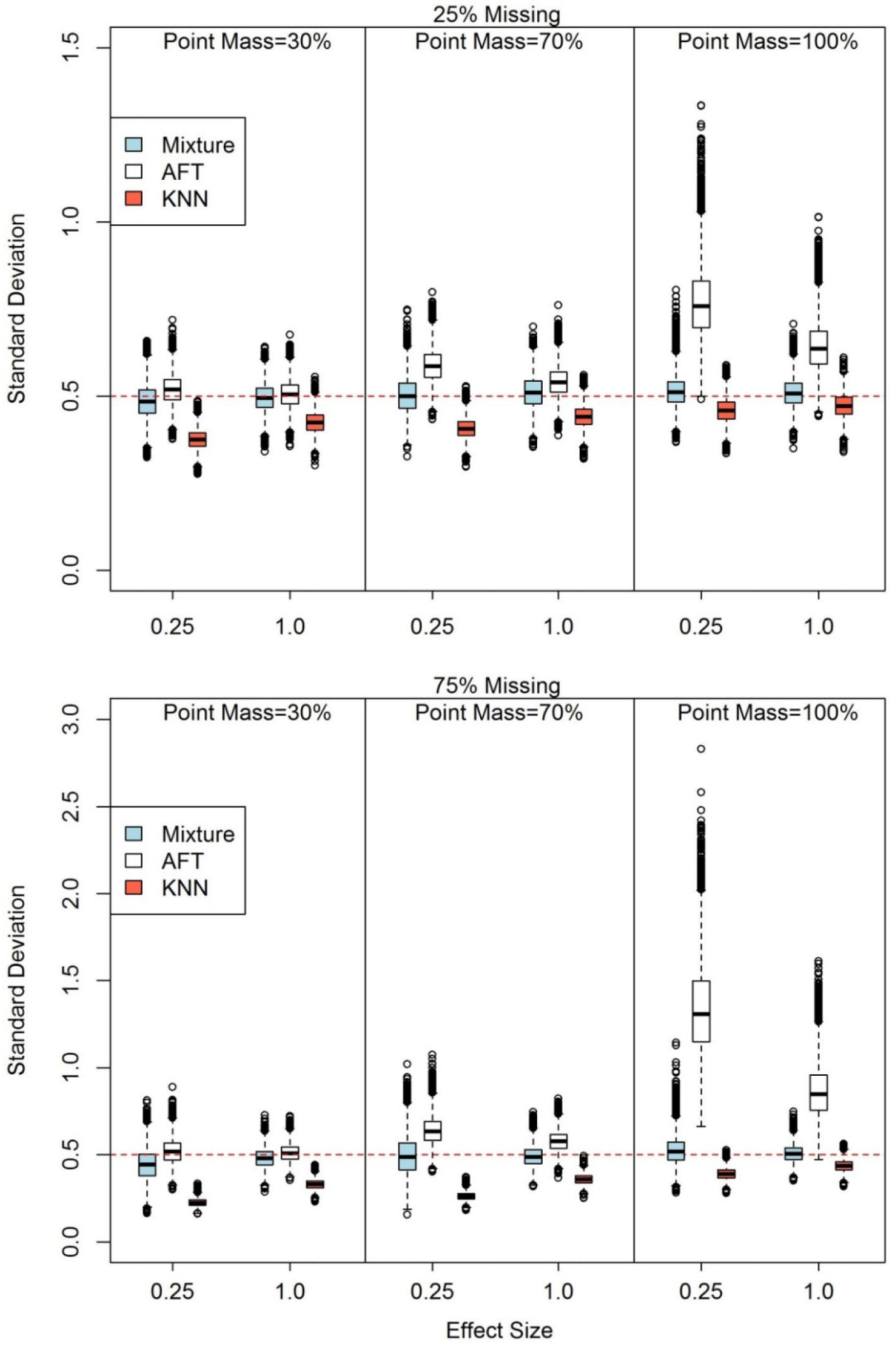

Figure 7.

Distribution of the standard deviation estimates based on the mixture and AFT models for 10,000 simulated data sets. Data were simulated from a log normal distribution with a mean of 13 for the Control group and means of 13.25, 13.5, and 14.0 for the Case group reflecting, effect sizes (i.e., Δ) of 0.25, 0.5, 1.0. The standard deviation was 0.5 for both the Control and Case groups. The proportion of missing values was 25 or 75% and the proportions of observations in point mass at 0 were 30%, 70%, or 100%. The sample size was 50 in each group. The horizontal red dashed line indicates the real value of the standard deviation used in simulating the data.

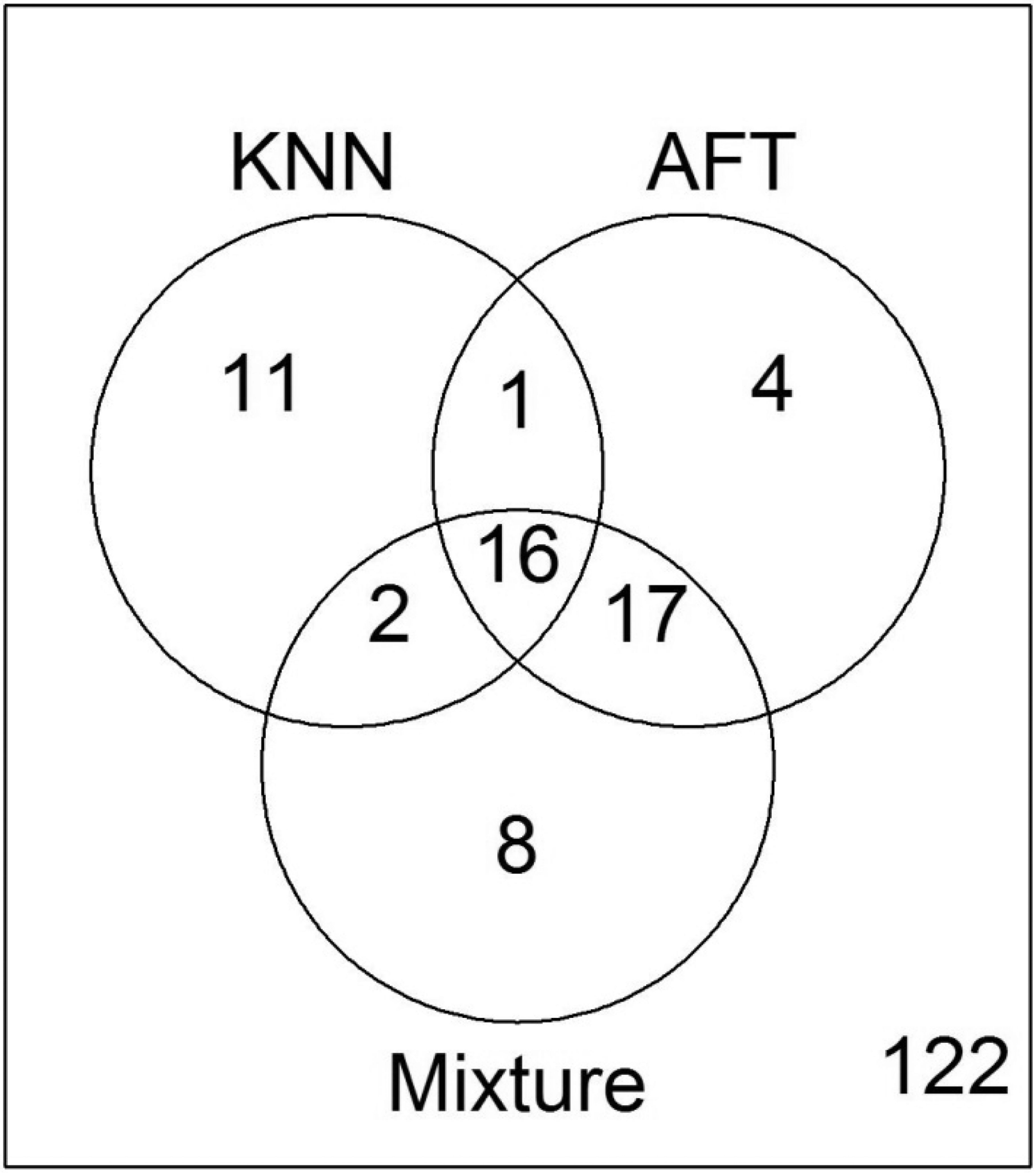

Figure 8.

Venn diagram of number of serum glycans found to differ significantly between women with ovarian cancer and healthy controls by with accelerated failure time model (AFT), mixture model (Mixture) and imputation with *k*nearest neighbors followed by a *t*-test (KNN). A total of 181 glycans were analyzed; 122 did not differ significantly with any method.

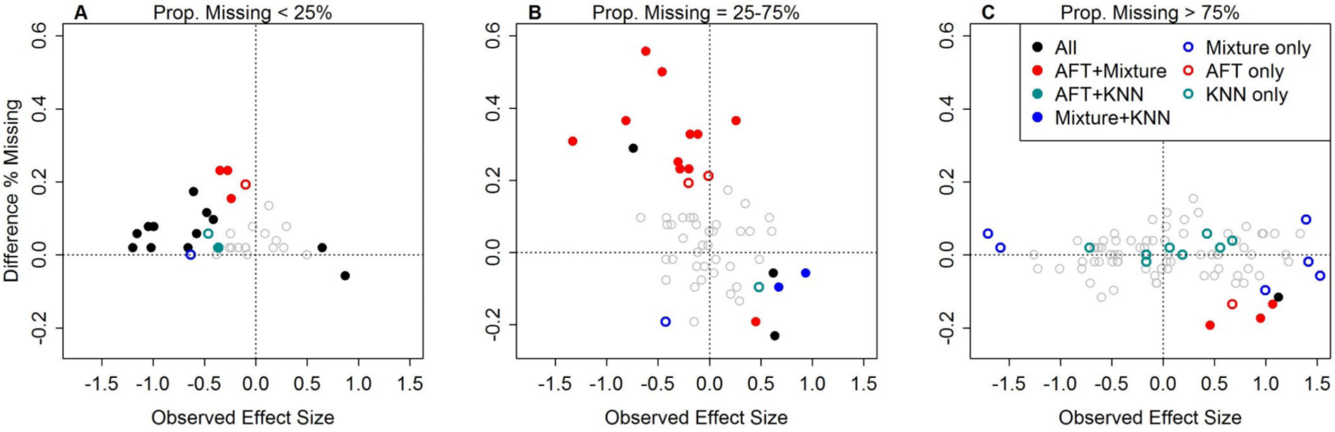

Figure 10.

Results of testing for differences in mean glycan values between women with ovarian cancer and healthy controls using the accelerated failure time (AFT) model, mixture model (Mixture), and imputation with *k*-nearest neighbors followed by at *t*-test (KNN). Gray circles indicate glycans that did not significantly differ (raw p-value ≥ 0.05) based on any test. Colored circles indicate glycans that differed significantly (raw p-value < 0.05) by cancer status based on at least one method. Results are plotted by the difference in the proportion of missing values between the two groups (y-axis) and the observed difference in means (x-axis) and are grouped according to the total proportion of missing values (cancer and controls combined). The observed effect size is the difference in the means of the observed values between cancer and control patients divided by the pooled standard deviation.

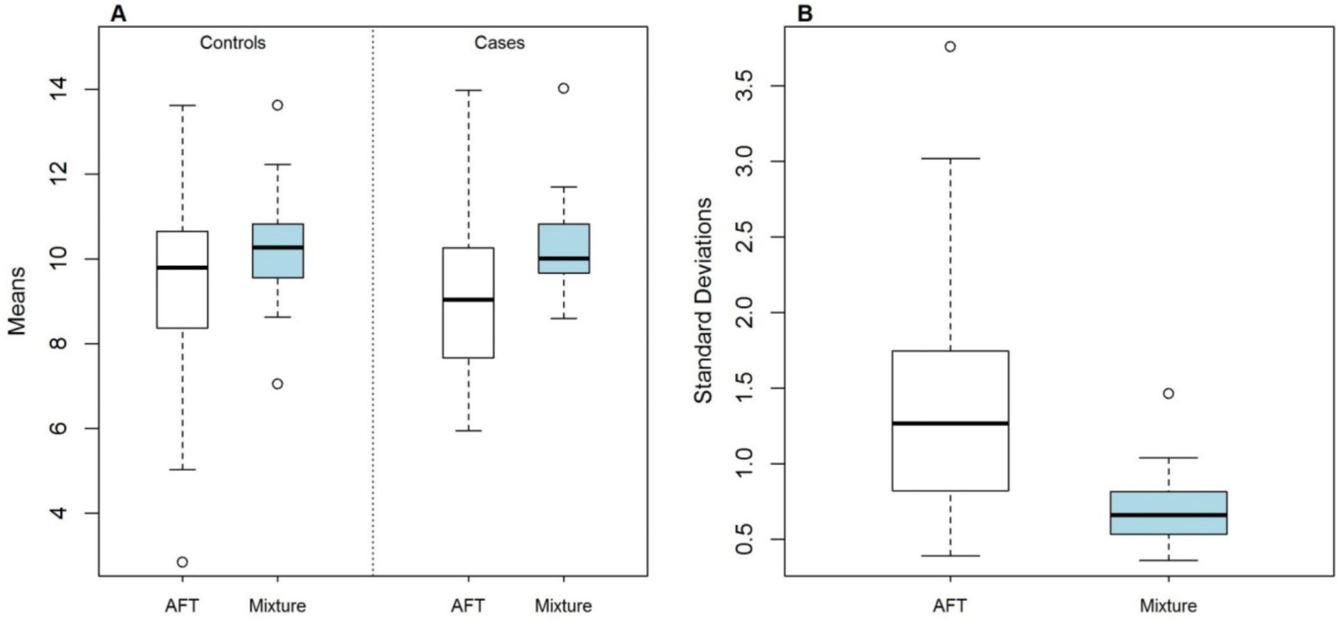

Figure 10.

Estimates of the A) means and B) standard deviation from AFT and mixture models for 38 glycans identified with the AFT model as significantly different between cancer and control subjects.

Table 1

Percentage of observations censored in the Case group for a given censoring level of the Control group for log normal distributions with standard deviations of 0.25, 0.50, and 1.

	Case Group	**Case Group**	**Case Group**
Control Group	**%Censored**	**%Censored**	**%Censored**
%Censored	**σ = 0.25**	**σ = 0.50**	**σ = 1**
25	4.7	12.0	17.8
50	15.9	30.8	40.1
75	37.2	56.9	66.4

Table 2

Type I error rate for the mixture and AFT models. Data were simulated from a log normal distribution with a mean of 13 and standard deviation of 0.5 for both the Control and Cases groups with n=50 and n=30 for both groups. The total Percentage of missing values was 25, 50, or 75 and the Percentage of missing values that were true zeros was 30, 70 and 100.

N=30	**% Missing**								
	25			**50**			**75**		
% Point Mass	**Mixture**	**AFT**	**KNN**	**Mixture**	**AFT**	**KNN**	**Mixture**	**AFT**	**KNN**
0 (Censored Only)	0.038	0.059	0.046	0.050	0.053	0.055	0.060	0.042	0.051
30	0.043	0.054	0.049	0.050	0.052	0.048	0.057	0.041	0.045
70	0.047	0.055	0.050	0.054	0.056	0.049	0.059	0.041	0.048
100 (No Censoring)	0.059	0.061	0.047	0.059	0.054	0.049	0.065	0.040	0.050

N=50	% Missing								
	25			50			75		
% Point Mass	Mixture	AFT	KNN	Mixture	AFT	KNN	Mixture	AFT	KNN
0 (Censored Only)	0.034	0.054	0.051	0.044	0.047	0.054	0.056	0.042	0.052
30	0.042	0.051	0.049	0.046	0.049	0.049	0.056	0.047	0.049
70	0.051	0.051	0.053	0.051	0.050	0.052	0.059	0.047	0.047
100 (No Censoring)	0.055	0.058	0.050	0.055	0.052	0.054	0.064	0.047	0.051